Research Report

Reporting Test Outcomes Using Models for Cognitive Diagnosis

Matthias von Davier

Lou DiBello

Kentaro Yamamoto

Research & Development

September 2006
RR-06-28

Reporting Test Outcomes Using Models for Cognitive Diagnosis

Matthias von Davier, Lou DiBello, & Kentaro Yamamoto
ETS, Princeton, NJ

September 2006

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

Models for cognitive diagnosis have been developed as an attempt to provide more than a single test score from item response data. Most approaches are based on a hypothesis that relates items to underlying skills. This relation takes the form of a design matrix that specifies for each cognitive item which skills are required to solve the item and which are not. This report outlines one direction that developments of cognitive diagnosis models is taking. It does not claim completeness, but describes a line of models that can be traced back to Tatsuoka's seminal work on the rule space methodology and that finds its current form in models that combine features of confirmatory latent factor analysis, multiple classification latent class models, and multidimensional item response models.

Key words: Cognitive diagnosis, skill profiles, multiple classification latent class models, item response models, general diagnostic model

Acknowledgments

The authors would like to thank Dan Eignor and two anonymous reviewers for their suggestions and comments on previous versions of the manuscript. The authors also thank Kim Fryer for editorial help. The opinions and conclusions contained in this paper are those of the authors and do not necessarily reflect the position or policy of ETS, or of the colleagues who reviewed the earlier version of the manuscript.

Models for Cognitive Diagnosis

The selection of models for cognitive diagnosis discussed here has been developed in an attempt to solve the diagnostic dilemma of both the classical and the probabilistic approaches to educational and psychological testing. Most models used for reporting student outcomes were originally developed to allow student behavior to be described using a single variable. Achievement, knowledge, and aptitude were thought of as essentially unidimensional, so a single number was deemed sufficient to describe them. This approach works when the purpose of the testing is to compare and eventually select students using a single criterion. This unidimensional view sees tests as tools to assign scores to certain fixed levels of achievement, rather than as tools to assess the current state in a process in which students are acquiring skills or knowledge.

Models for cognitive diagnosis are based on a different assumption, namely that observed differences in student performance on a set of tasks, even if correlated across tasks, are best described by more than one student attribute or skill, and that a multivariate profile is necessary to describe differences among examinees. This view sees tests as tools for better understanding and evaluating areas where there is potential for improvement. Over the years, different approaches have been taken to formalize these different sets of assumptions:

The most common models of modern test theory for the univariate student model are item response theory (IRT; Lord & Novick, 1968) and the Rasch model (RM; Rasch, 1960). These models are being used operationally in many K-12 testing programs, as well as in large-scale educational survey assessments that report on representative samples of student populations in different grades, states, or even countries (von Davier, Sinharay, Oranje, & Beaton, 2006).

Describing student behavior on a set of cognitive tasks assuming multiple discrete skills or attributes has been an active area of research for quite some time. The rule space methodology (Tatsuoka, 1983) and latent class models with multivariate latent spaces (Haberman, 1979; Haertel, 1989; Maris; 1999) are the most well-known early attempts at diagnostic modeling. More recent approaches are the unified model (DiBello, Stout, & Roussos, 1995) and the reparameterized unified model (RUM), also referred to as the fusion model, (Hartz, Roussos & Stout, 2002), as well as approaches that involve Bayesian networks. Recently, a class of models referred to as the general diagnostic model (GDM; von Davier & Yamamoto, 2004a, 2004b; von Davier, 2005) has been developed. It has been shown that this class of models contains many of the previous approaches, in addition to some common IRT models as special cases.

Most diagnostic models assume a multivariate but discrete latent variable which represents the absence or presence of multiple skills or attributes. These skill profiles have to be inferred through model assumptions with respect to how the observed data of an examinee relate back to the unobserved skill profile. The absence or presence of skills is commonly represented by a Bernoulli (0/1) random variable in the model. Given that the number of skills represented in the model is larger than in unidimensional models (obviously greater than 2, but smaller than, say, 14 skills in most cases), the latent distribution of skill profiles needs some specification of how to model the relationship between skills in order to avoid the estimation of up to 2^{14}-1 = 16,383 separate skill pattern probabilities. The GDM (von Davier, 2005) allows ordinal skill levels and different forms of skill dependencies to be specified so that more gradual differences between examinees can be modeled in this framework.

Rule Space Methodology

Rule space refers to a two-dimensional representation of an unidimensional ability and a deviation from a model based on the unidimensional ability variable. More specifically, the rule space methodology uses ability estimates from IRT and discrepancy scores from person-fit measures to span a two-dimensional space of ability and deviation from the IRT model. Assume there are $x_1,...,x_I$ observed dichotomous (0/1) responses to I items or tasks for each of N examinees in a sample. For the IRT model used in rule space, we may assume that the probability of observed response vector $(x_{1n},...,x_{In})$ for examinee n is given by

$$P(x_{1n},...,x_{In}) = \prod_{i=1}^{I} P(x_{in}|\theta_n, \beta_i) \qquad (1)$$

with item responses following the 3PL model, that is,

$$P(x=1|\theta, \beta=(a,b,c)) = c + (1-c)\frac{1}{1+\exp[-a(\theta-b)]}.$$

The ability estimate θ of an examinee is the value that maximizes Equation 1 given item parameters β and depends on the responses $(x_{1n},...,x_{In})$.

Person-fit measures are used to identify examinees who exhibit unexpected response patterns, that is, response patterns with a low probability under the assumed model (see von

Davier & Molenaar, 2003). One person-fit measure that may be used in the rule space methodology is defined as

$$\xi_n = 1 - \frac{\mathrm{cov}(x_{ni}, x_{.i})}{\mathrm{cov}(x^s_{gi}, x_{.i})} \qquad (2)$$

where the covariances are defined across items and are computed based on the covariance between item responses of examinee *n* and the column sum (item sum) of success for the numerator. For the denominator, the covariance (across items) of the Guttman pattern for a given sum score *s* and these column sums are computed. This defines a measure of person fit that is independent of the overall ability estimate θ (Tatsuoka, 1983) and indicates how deviant a specific response pattern is from the ideal (Guttmann) pattern for the given examinee.

Using θ and ξ for each observed response pattern allows a two-dimensional scatter plot of the examinees to be created. This first building block is complemented with an expert generated matrix that relates items to the skills required to solve those items. This matrix is referred to as the Q-matrix. The Q-matrix contains zeros and ones. The nonzero entries relate the cognitive tasks (items) to a set of skills that is assumed to drive student responses on these tasks. If an examinee has all the skills required to complete a particular task, his or her probability of success on this task should be high. If, however, an examinee lacks certain skills required to complete a particular task, the success expectation should be low for that task.

Table 1 gives an example of such a Q-matrix for two examinees (examinee y and examinee z) and their respective skill sets. Assume that for a given math assessment, skills to add, subtract, and multiply are required, denoted by Add., Sub, and Mult in the table. If an examinee has all the skills required to complete a particular test item, the probability that the examinee will solve the item is high. As shown in Table 1, examinee y lacks skills Add and Sub., but interestingly has the skill to carry out multiplications. This implies that the probability of success for student y reaches its maximum only for item F, since this is the one item that requires the Mult. skill only, and no additional skills. In contrast, Table 1 indicates that examinee z will solve all items A to F with comparably high probability, since he or she possesses all three skills required by this set of items.

The implied rule when converting a Q-matrix and a skill pattern to a set of expected responses is: the more required skills present, the higher the probability of success. This assists

in determining the most probable responses for each set of skills. For examinee y in the example one may argue that (A = 0, B = 0, C = 0, D = 0, E = 0, F = 1, G = 0) is the most plausible vector of responses if the presence of all required skills is necessary to solve a specific task. This view would represent a *noncompensatory* approach underlying the way in which skills are expressed or translated into success rates. A somewhat more forgiving view could be that an examinee may either show the above pattern of responses or produce at least one other response pattern, namely (A = 0, B = 1, C = 0, D = 0, E = 1, F = 1, G = 1), since at least a fraction of the required skills are present. This represents a *compensatory* assumption of how skill presence is expressed in higher or lower probabilities of succeeding in tasks. For examinee z, however, all required skills are present, so the typical response from this examinee should be (A = 1, B = 1, C = 1, D = 1, E = 1, F = 1, G = 1).

Table 1

Fictitious Q-Matrix for Six Items, Three Skills, and Two Examinees y and z With Different Skill Sets

	Q-matrix: task by skill			Examinee y			Examinee z		
Skill	Add.	Sub.	Mult.	Add.	Sub.	Mult.	Add.	Sub.	Mult.
Task				no	no	yes	yes	yes	yes
A	1			-			+		
B	1		1	-		+	+		+
C	1	1		-	-		+	+	
D		1			-			+	
E		1	1		-	+		+	+
F			1			+			+
G	I	I	I	-	-	+	+	+	+

Note. Add, =addition; Sub, = subtraction; Mult = multiplication.

In this example, there are eight different skill profiles, since all three skills can be either absent or present. These eight profiles correspond to eight typical response patterns as illustrated above. Table 2 illustrates these ideal (or perhaps, most typical) response patterns of the eight different skill profiles, assuming a noncompensatory model.

Table 2

Three Skills and Their Associated Typical Response Patterns Under Noncompensatory Assumptions for The Example Q-matrix From Table 1

	Skills			Tasks					
S	Add.	Sub.	Mult.	A	B	C	D	E	F
000	no	no	no	0	0	0	0	0	0
100	yes	no	no	1	0	0	0	0	0
010	no	yes	no	0	0	0	1	0	0
110	yes	yes	no	1	0	1	1	0	0
y	no	no	yes	0	0	0	0	0	1
101	yes	no	yes	1	1	0	0	0	1
011	no	yes	yes	0	0	0	1	1	1
z	yes	yes	yes	1	1	1	1	1	1

Note. Add. =addition; Sub. = subtraction; Mult. = multiplication.

The typical item response patterns contain a 1 whenever all required skills in the Q-matrix are present in the skill profile. The above eight response patterns are used in the rule space methodology to define the centroids of clusters of examinees that are close to these patterns (and finally, assumed to belong to the associated skill profile group).

More explicitly, each expected item response pattern has a corresponding value on the skill profile. In addition, each expected response pattern, for example $\overrightarrow{x_{110}} = (1,0,1,1,0,0)$, also corresponds to a specific point in the ability-by-item fit plane, that is, in the example S=110, there is a $(\theta,\xi)_{110}$ that represents the skill pattern 110 (Add. = yes, Sub. = yes, Mult. = no) in the space spanned by ability and item fit measure, θ and ξ.

In our example, the rule space methodology would determine these (eight) points, $(\theta,\xi)_{000}$ to $(\theta,\xi)_{111}$, which are the centroid points corresponding to typical responses of different skill patterns in the ability-by-person fit space. Then, the rule space method would classify each examinee in the sample, based on his or her observed item response pattern, $\vec{x} = (x_1,...,x_6)$ into one of the clusters defined by these (eight) centroids. The building blocks of the two-dimensional space used in rule space are IRT and person-fit measures. Alternately, rule

space assumes that items are solved using multiple skills, which are required to a different extent in different items. This implies the bringing together of a unidimensional IRT model with the use of multiple skills to solve the items. One issue is the connection between response pattern, underlying ability, and skill pattern, and how these three parts can be combined with classification rules that use the ability estimate and a person-fit measure. This classification can be done using different deterministic or probabilistic classifiers or rules, depending on the preference of the user of the methodology.

The questions to be answered are: If deterministic classifiers are chosen, what happens if some response patterns do not have very pronounced proximity to a unique skill pattern? Which measure of proximity to a centroid is used to classify observations? If probabilistic classifications are used, what density or model is assumed to calculate the probability of an examinee being a member of a skill-pattern class or cluster given an observed response pattern?

Subsequently developed models for cognitive diagnosis embed the skill-by-item Q-matrix more directly into the model structure, instead of resolving the duality of skills and IRT-based ability estimates in the classification phase. The models described in the following sections start out with probabilities explicitly modeled based on multiple skills and an expert-generated Q-matrix, or they provide means to incorporate multiple latent dichotomies, which may be viewed as multiple mastery/nonmastery groups.

Multiple Classification Latent Class Models

Latent class analysis (LCA) assumes a categorical latent variable that explains the observed relationships between examinees' item responses. The defining properties of LCA are local independence given latent class, the assumption of an exhaustive and disjunctive latent classification variable, and distinctness of conditional probabilities across classes. Why is that a useful approach to cognitive diagnosis modeling?

The local independence assumption ensures that, given latent class, the observed variables (responses to cognitive tasks or items) are independent. This means that examinees belonging to the same latent class differ from the ideal class profile in their responses only randomly; they do not show any further systematic variation in their item responses.

The assumption of the latent classes being disjunctive and exhaustive ensures that each examinee is a member of exactly one latent class. This means that each examinee, given

sufficient item response information, can be classified into one of the latent classes with high probability if the model holds.

Latent classes differ with respect to the profile of response probabilities across items. It is assumed that each class is defined by a unique set of probabilities that, in many ways, defines the particular class.

The model equation, which represents the formal structure that corresponds to these assumptions, is

$$P(x_1,...,x_I) = \sum_{c=1}^{C} \pi_c \prod_{i=1}^{I} P(X = x_i | c, i), \quad (3)$$

where the π_c are the relative class sizes, and the $P(X = x_i | c, i)$ are conditional probabilities of a response *x* on item *i* in class *c*. If the classification variable is latent; that is, the membership of the examinees is unobserved, the conditional probabilities and the relative sizes of the classes have to be estimated from the data.

One may view the LCA, even in this unrestricted form, as a model for cognitive diagnosis. If all parameters are freely estimated, the LCA is an exploratory model. In this form the model can provide useful insight into how observed responses may be represented as being based on an unobserved mixture of different (cognitive) types or classes of individuals. If there are hypotheses about specific expected profiles that correspond to different cognitive styles or classes, however, LCA can be used to analyze data and directly incorporate these hypotheses into the conditional probabilities. Haertel (1989) and Maris (1999) provide a formal introduction and examples of how to execute such an analysis using LCA.

Instead of paraphrasing Haertel's and Maris's work, we will introduce the way to find parameter constraints for LCA that relate closely to the rule space methodology. Methods to implement constrained latent class analysis can be used to develop a method to directly incorporate assumptions about skill expression and an expert-generated Q-matrix into the conditional probabilities of LCA. Table 3 provides an example of how to specify latent class probabilities for the two examinees from the example in the previous section.

The conditional probabilities are constituted to be equal for all items through skill-pattern combinations where the number of present skills and the number of required skills coincide. As an example, if a person with a specific skill pattern has both of two required skills, his or her

conditional probability will be *P(2/2)*; if a person has only one of two required skills, his or her conditional probability will be *P(1/2)* for all the items for which this ratio holds. Recall the rule space example where it was argued that examinee y has the highest probability for item F since this examinee has the Mult. skill, which is the only one required for that item.

Table 3

Fictitious Conditional Success Probabilities to Incorporate the Q-matrix Assumptions and Assumed Skill Patterns Into a LCA Type Analysis

Q-matrix: task by skill					
Task	Add.	Sub.	Mult.	y = 001	z = 111
A	1			P(0/1)	P(1/1)
B	1		1	P(1/2)	P(2/2)
C	1	1		P(0/2)	P(2/2)
D		1		P(0/1)	P(1/1)
E		1	1	P(1/2)	P(2/2)
F			1	P(1/1)	P(1/1)
G	1		1	P(1/2)	P(2/2)

Note. Add., =addition; Sub., = subtraction; Mult. = multiplication.

The above approach generates a vector of conditional probabilities for all skill patterns (000 to 111 in our example). The corresponding restricted LCA model for cognitive diagnosis contains $8 = 2^3$ latent classes for the example with three dichotomous skill variables, but instead of 8x7=56 conditional probabilities, only five different probabilities, *P(0/1)*, *P(1/1)*, *P(0/2)*, *P(1/2)*, and *P(2/2)*, have to be estimated.

The uniqueness of the class-specific profiles is one of the issues that should be monitored when using a restrictive latent class model. Each class (read: skill pattern) profile of conditional probabilities consists only of these probabilities, so skill patterns that are similar in certain ways may produce very similar patterns of conditional probabilities (read: class profiles). The restrictions of this model are, simultaneously, a source of strength when using this parameterization suggested by Yamamoto (1992). If the number of skills is large compared to the number of items, very large numbers of parameters have to be estimated based on a moderate set of item responses per examinee. In the restricted cognitive diagnosis LCA, this potentially

huge number of parameters is cut down to very few levels of conditional probabilities. The number of probability levels to be estimated depends on the maximum number of skills required per item. In addition, the number of latent class sizes is determined by the number of skill levels to the number of skills; in our example this is $8 = 2^3$. This number increases exponentially with the number of skills involved; for example, when using 10 skills, the number of possible skill patterns increases to 1,024.[1]

Reparameterized Unified Model

This section presents a brief description of the reparameterized unified or fusion model (adapted from DiBello & Stout, 2003). In addition, we present two ways to parameterize the fusion model to allow estimation with partial credit items and with skills that contain an arbitrary finite number of ordered levels.

Probabilistic Unified Model—Dichotomous Items and Skills Case

The unified model was developed as a probabilistic item model to express the stochastic relationship between item response and status of underlying skills (DiBello, Stout, & Roussos, 1995). A thorough discussion of the model is beyond the scope of this paper, but we present just enough to give an idea of the modeling so it can be compared to other models presented in this chapter.

As a starting point we assume an underlying conjunctive latent response model as employed within the rule space method (Tatsuoka, 1985, 1990, 1995) and later within the latent response models of Maris (1999). We then select a moderately sized set of skills that are important to the client and that we believe are able to be well measured by a test. Settling on an appropriate list of skills at the right level of granularity is a critically important step (VanEssen, 2001). The primary assumption that underlies the fusion model is conjunctive; sufficient proficiency in all of the indicated skills is required to successfully answer the item.

A student's proficiency is modeled as an unobservable profile of skill level $\underline{\alpha} = (\alpha_1, ..., \alpha_K)$. Here, $\alpha_k = 1$ means that skill *k* is mastered and $\alpha_k = 0$ means that skill *k* is not mastered. The relationship between item response and skill is modeled as conjunctive (noisy AND model in the sense of Junker and Sijtsma, 2001) in that an item is considered to involve a certain subset of *k* skills, and an imaginary deterministic response to that item for that student would be correct if, and only if, the student has mastered *all* the skills required by that item. If one or more of those

required skills is not mastered, the deterministic response would be incorrect. To derive the unified model, we begin with a latent variable characterization of the cognitive assumptions underlying the unified model. The item response $X_i = 0/1$ is expressed as

$$X_i = S_i \left[\prod_{k=1}^{K} (Y_{i,k})^{q_{i,k}} \right] B_i + (1 - S_i) C_i$$

where S_i, $Y_{i,k}$, B_i, and C_i are dichotomous variables with values *0/1* considered to be latent response variables in the sense of Maris (1999) with the following definitions:

S_i= examinee chooses the Q strategy for item i (*1*=yes; *0*=no);

$Y_{i,k}$= skill k is applied correctly to item i (*1*=yes; *0*=no);

B_i= other knowledge or skills not included in Q are performed correctly for item i (*1*=yes; *0*=no);

C_i= given that a different strategy is used for item i, that strategy is used correctly (*1* = yes; *0* = no).

It is not claimed that students behave deterministically, or that they go through these various stages in a conscious or orderly fashion. This latent response model only provides a convenient way to list cognitive aspects that govern response performance by students on items or tasks. The point of this characterization is twofold: first, the model focuses in greatest detail on the particular skills listed in the Q-matrix; second, there are other things going on outside the Q-matrix that may be important enough to attend to.

Consequently, the fusion model includes a probabilistic element to represent stochastic variation from the deterministic response. A student's observed response is thought of as a noisy version of the deterministic latent response. The most general form of the unified model identifies four specific factors to explain the divergence of observed from latent response on an item.

Strategy. For each item the Q-matrix presumes a predominant solution strategy and specifies a particular set of skills that are required by that strategy for successfully answering the item. In general, other strategies requiring different sets of skills may be available to solve this problem, and the student may choose, consciously or not, to employ a strategy different from that

embodied in the Q-matrix. We introduce parameter d_i = probability across the population that the Q strategy is selected for item *i*.

Completeness. The Q-matrix represents a manageable set of skills available to be coded as required for individual items. For various reasons, we may decide to leave out of the Q-matrix some skills that are important for a particular item. For example, Samejima (1995) suggested leaving out higher order thinking skills from the Q-matrix, since these are difficult to discretize. Obviously, skills of too fine granularity (i.e., skills that are only required by very few items) should be left out or combined into super-skills to allow reliable identification of mastery/nonmastery of these fewer skills. In some situations the skill set represented in the Q-matrix does not completely represent everything necessary for successfully answering an item. In those cases the Q-matrix is incomplete for that item. For a given item *i*, a parameter c_i is used as an index of the extent to which the Q-matrix is complete for that item. We introduce a continuous student ability parameter η to represent overall ability outside the skills modeled in the Q-matrix. In contrast to the detailed modeling of the skills listed in the Q-matrix, abilities outside the Q-matrix are modeled as simply as possible. We do not intend to imply that the non-Q abilities are less important, but we consciously focus on the cognitive aspects that are modeled more explicitly in the Q-matrix. The non-Q abilities are crudely lumped together and modeled by a continuous variable η. Often, η is assumed to be unidimensional, and it acts as a variable that collects variation outside that explained by the skills listed in Q.

Positivity. A person may be a master of skill *k* and yet apply the skill incorrectly to item *i*. For example, a specific instance of a skill within an item is particularly challenging so that even masters of the skill may occasionally misapply it in that particular item. Conversely, a nonmaster of skill *k* may correctly apply the skill to a particular item because the item instance is easy. For example, a nonmaster of a fractions skill may be able to correctly handle very simple fractions (such as ½) in a problem coded as requiring fractions knowledge, because the student knows enough about fractions to handle ½ correctly, even if his or her overall fractions proficiency is not high enough to render the student a master of that fractions skill. Modeling this formally requires two parameters for each item-skill pair.

$\pi_{i,k}$ = P(apply skill *k* correctly in item *i*|skill *k* is mastered)

$r_{i,k}$ = P(apply skill *k* correctly in item *i*|skill *k* is not mastered)

Slips. Given that everything else has been done correctly, a response that should have been correct may be recorded incorrectly. We call this a slip and introduce the slip parameter *p* = probability of committing a slip.

The unified model assigns parameters for each of these factors and provides a parametric expression for the probability of a correct response to an item, given a student's mastery/nonmastery skill pattern.

$$P(X_i = 1 | \underline{\alpha} = (\alpha_1, ..., \alpha_K), \eta) =$$
$$(1-p)\left\{ d_i \left[\prod_{k=1}^{K} (\pi_{i,k})^{q_{i,k}\alpha_k} (r_{i,k})^{q_{i,k}(1-\alpha_k)} \right] P_{c_i}(\eta) + (1-d_i) P_{b_i}(\eta) \right\}$$

Here η = continuous ability outside of the Q skills.

$P_b(\eta) = 1 \big/ (1 + \exp(-1.7(\eta - b))$ = one parameter logistic with *a=1* and *1.7* present.

b_i = "difficulty" parameter for non-Q strategy.

$0 \le c_i \le 3$ is the item completeness index.

$c_i \approx 0$ implies that other (unspecified) skills are important for answering item *i* correctly.

$c_i \approx 3$ implies that the specified attributes in Q suffice to explain examinee response to *i.*

In moving from the latent response model to this item model, several conditional independence assumptions are made. For further details see DiBello, Stout, and Roussos (1995) and Hartz (2002).

Features of the Unified Model

The unified model has several important characteristics. Rather than continuous IRT θ, the latent space now is a mixed discrete-continuous $(\underline{\alpha}, \eta)$ variable. The skill profile vector $\underline{\alpha} = (\alpha_1, ..., \alpha_K)$ represents mastery or nonmastery for each of the skills being diagnosed by the test. The continuous parameter η represents ability outside the skills listed in the Q-matrix. The four aspects of the model — strategy, completeness, positivity, and slips — are designed to capture more of the reality of the examinee response process where the underlying conjunctive cognitive

model seems reasonable. Such a model should be more close to reality and lead to more accurate and valid diagnoses.

One may argue that the model is still far too simple, that it is not consistent with dynamic aspects of cognition, and that it does not express the rich interplay between item characteristics and cognitive functioning. All of these arguments are correct, but the unified model is tractable. Statistical approaches inherently cannot deal with the full complexities of cognitive processing in response to assessment tasks. We believe the unified model may be useful for formative assessments.

Fusion Model

The original unified model lacked a practical calibration method, and the $\pi_{i,k}$ s and $r_{i,k}$ s were nonidentifiable. In her PhD thesis, Hartz (2002) reparameterized the model, cast it into a hierarchical Bayesian framework, and programmed a Markov chain Monte Carlo (MCMC) parameter estimation procedure called Arpeggio. Hartz chose a particularly intuitive and useful reparameterization:

$$\pi^*_i = \prod_{k=1}^{K} (\pi_{i,k})^{q_{ik}} \text{ and } r^*_{i,k} = \frac{r_{i,k}}{\pi_{i,k}}.$$

These K_i+1 *-parameters are statistically identifiable given that sufficient information is available from the item-response data matrix.

The *-parameters also lend themselves to interpretation by nontechnical test users and test developers:

π^*_i can be thought of as a conditional item difficulty parameter, conditional on having mastered all the skills required by an item;

$r^*_{i,k} = \frac{r_{i,k}}{\pi_{i,k}}$ is a certain weight of evidence and measures the inverse information strength of skill *k* within item *i*. If $r^*_{i,k}$ is low, near *0.0*, the penalty to probability of correct item response for nonmastery of skill *k* is high. If $r^*_{i,k}$ is high, near *1.0*, the penalty is low.

Once a test is calibrated, provided model-data fit is high, the values $r^*_{i,k}$ quantify how well skill *k* is tapped by item *i.* Reliability of classification of skill *k* by the test as a whole

depends on the item parameters and correlations among the skills. The fusion model provides a psychometrically sound method to empirically support expert-judgment-based standards or alignments of skills to items. Employing Hartz's reparameterization gives the RUM

$$P(X_i = 1 \mid \underline{\alpha} = (\alpha_1, ..., \alpha_K), \eta) = \left[\pi_i^* \prod_{k=1}^{K} (r_{i,k}^*)^{q_{i,k}(1-\alpha_k)} \right] P_0(\eta + c_i)$$

The term fusion model refers to the RUM cast in a hierarchical Bayesian framework (see Hartz, 2002; Stout & Hartz, 2004).

Using the Fusion Model with Partial Credit Items

Suppose item *i* has ordered scores $X_i = 0,1,...,M_i$ where M_i is the maximum possible score and $M_i + 1$ is the number of possible score levels. A straightforward way to use the fusion model for partial credit scored items is as follows (see Bolt & Fu, 2004):

$$P_{im}^*(\underline{\alpha}, \eta) = P(X_i \geq m \mid \underline{\alpha} = (\alpha_1, ..., \alpha_K), \eta) = \begin{cases} 1 & m = 0 \\ \left[\pi_{i,m}^* \prod_{k=1}^{K} (r_{i,k,m}^*)^{q_{i,k}(1-\alpha_k)} \right] P_{c_{i,m}}(\eta) & m = 1, 2, ..M_i \end{cases}.$$

From this we define item score probabilities as follows:

$$P_{i,m}(\underline{\alpha}, \eta) = P(X_i = m \mid \underline{\alpha} = (\alpha_1, ..., \alpha_K), \eta) = \begin{cases} P_{i,m}^*(\underline{\alpha}, \eta) - P_{i,m+1}^*(\underline{\alpha}, \eta) & m = 0, ..., M_i - 1 \\ P_{i,M_i}^*(\underline{\alpha}, \eta) & m = M_i \end{cases}.$$

Where $\pi_{i,m}^*$ is the probability of applying skill *k* well enough in item *i* to achieve an item score of at least *m* given all required skills are mastered ($\pi_{i,1}^* \geq \pi_{i,2}^* \geq ... \geq \pi_{i,M_i}^*$)and $r_{i,k,m}^*$ is the ratio of two probabilities: 1) the probability of applying skill *k* in item *i* well enough to achieve an item score of at least *m* given the examinee is a nonmaster of skill *k*, and 2) the probability of applying skill *k* in item *i* well enough to achieve an item score of at least *m* given the examinee is a master of skill *k* ($r_{i,1}^* \geq r_{i,2}^* \geq ... \geq r_{i,M_i}^*$) and $P_{c_{i,m}}(\eta)$ is a one parameter logistic function with completeness parameter $c_{i,m}$, that represents the probability of performing all non-Q knowledge well enough to achieve item score of at least *m* ($c_{i,1} > c_{i,2} > ... > c_{i,M_i}$).

In the case of a dichotomous item score ($M_i = 1$) this reduces to the dichotomous fusion model. In fact, as Bolt and Fu (2004) point out, this representation amounts to assuming a dichotomous fusion model for each possible dichotomization of the partial credit item scores

$$X_i = 0,1,...,M_i .$$

The number of parameters for this version of the fusion model is quite high. In general, the number of parameters for an item with maximum score M_i that requires K_i skills is

$$M_i(K_i + 2) .$$

Bolt and Fu propose an alternate parameterization that reduces the total number of parameters by imposing constraints on the item parameters. The reduced parameterization is seen most easily in the case of a dichotomous item that requires only one skill *k*. For convenience we suppress the subscript *k.* Imagine an underlying normal propensity variable t_i that represents the propensity for applying all required skills correctly in this item (in this example, there is only one skill required).

Then we consider the probability of correctly applying all skills (in this case one skill) correctly in the item to be represented by the area under the standard normal curve to the right of a threshold value $\tau_{i,1}$.

Next, imagine the item is partial credit with maximum score M_i. We can consider the same underlying normal propensity variable t_i. The application of all the required skills (in this case the one skill *k*) well enough to achieve item score $X_i \geq m$ for each score level *m* is represented by M_i thresholds on the standard normal curve $\tau_{i,1} \leq \tau_{i,2} \leq ... \leq \tau_{i,M_i}$.

Next, consider a partial credit item with maximum score $M_i = 2$ that requires three skills $k = 1,2,3$. Now instead of one, we consider four *(K+1)* underlying normal propensity variates: $t_{i,1}, t_{i,2}, t_{i,3}, t_{i,4}$. Each of these normal distributions represents a different conditional underlying propensity for applying all three required skills correctly in the item. The four curves are conditional on the four attribute states: (111), (011), (101), and (110).

Bolt and Fu (2004) make the simplified assumption that each of these normal curves has the same variance *1.0*, and the curves differ only with respect to the location of the mean. The first curve (111) is located at mean *0.0*, and the next three curves at the following means: $\mu_{i,011} = \mu_{i,1}$, $\mu_{i,101} = \mu_{i,2}$, $\mu_{i,110} = \mu_{i,3}$. These curves are thought of as existing all on the same underlying scale, and the means are constrained to be located to the left of *0.0*.

If we scale each of these four underlying propensities to lie on the same scale, then the thresholds $\tau_{i,1} \leq \tau_{i,2} \leq ... \leq \tau_{i,M_i}$ are the same for all four curves. Thus, we represent the $M_i = 2$ item score levels by locating $M_i = 2$ thresholds $\tau_{i,1} \leq \tau_{i,2}$ and think of these thresholds as holding for all four normal curves. The conditional probability that all three required skills are performed well enough to score at least item score *m*, conditional on one of the four skill patterns (111, 011, 101, or 110), is represented by the area under the relevant curve (111, 011, 101, or 110) to the right of the threshold $\tau_{i,m}$.

It is easy to see that the parameters $\pi^*_{i,m}$ and $r^*_{i,k,m}$ can be defined in terms of the parameters $\mu_{i,k}$ and $\tau_{i,m}$. An MCMC estimation program has been developed and evaluated for estimating the $\mu_{i,k}$ and $\tau_{i,m}$ (Bolt & Fu, 2004).

The advantage of this parameterization is a significant reduction in the number of parameters. Instead of $M_i(K_i + 1)$ parameters $\pi^*_{i,m}$ and $r^*_{i,k,m}$, the new parameterization requires only $M_i + K_i$ parameters $\mu_{i,k}$ and $\tau_{i,m}$ parameters. It should be noted that the $\mu_{i,k}$ and $\tau_{i,m}$ parameterization is not equivalent to the original $\pi^*_{i,m}$ and $r^*_{i,k,m}$ parameterization. Making the variances of the *K+1* curves equal is a restrictive assumption. The parameters $\pi^*_{i,m}$ and $r^*_{i,k,m}$ can be derived from the parameters $\mu_{i,k}$ and $\tau_{i,m}$. The converse requires constraints on the $\pi^*_{i,m}$ and $r^*_{i,k,m}$. Bolt and Fu (2004) argue that the constraints are reasonable for educational testing applications. For further details, see Bolt and Fu.

Extension of the Fusion Model to Ordered Skill Levels

Templin (2004) extended the fusion model to the case in which each skill *k* can have an arbitrary number of ordered levels: $\alpha_k = 0,1,...,l_k$. Here we describe only the case of dichotomous items with polytomous skills. This can be combined with the Bolt-Fu partial credit

approach above, which has been done, and both real and simulated data studies have been performed (Templin, 2004; Templin, Roussos, & Stout, 2004).

Consider a dichotomous item *i* that requires a number of skills *k=1,2,3*. Each skill has an arbitrary number of levels $\alpha_k = 0,1,...,l_k$. The ordered, polytomous fusion model is defined by Templin as

$$P(X_i = 1 | \underline{\alpha} = (\alpha_1,...,\alpha_K), \eta) = \left[\pi_i^* \prod_{k=1}^{K} (r_{i,k}^*)^{f_{i,k}(\alpha_k, q_{i,k})} \right] P_{c_i}(\eta).$$

Here the linking functions $f_{i,k}(\alpha_k, q_{i,k})$ satisfy the following constraints:

1. $f_{i,k}(\alpha_k = 0, q_{i,k} = 1) = 1$
2. $f_{i,k}(\alpha_k = l_i, q_{i,k} = 1) = 0$
3. $f_{i,k}(\alpha_k = m, q_{i,k} = 1) > f_{i,k}(\alpha_k = m+1, q_{i,k} = 1)$ for *m=0,1,2, l-1*.

Templin (2004) developed MCMC software to calibrate this model and has performed several research studies of real test data as well as parameter recovery studies with simulated data. Templin notes that the ordered polytomous fusion model is equivalent to replacing the single polytomous skill $\alpha_k = 0,1,...,l_k$ with l_k dichotomous subskills obeying an order constraint. For example we can define

$$\alpha_{k\_1} = \begin{cases} 0 \; if \;\; \alpha_k = 0,1,...,l_i - 1 \\ 1 \qquad\quad if \;\; \alpha_k = l_i \end{cases}$$

$$\alpha_{k\_2} = \begin{cases} 0 \; if \;\; \alpha_k = 0,1,...,l_i - 2 \\ 1 \qquad if \;\; \alpha_k = l_i - 1, l_i \end{cases} \wedge$$

…

$$\alpha_{k\_l_i} = \begin{cases} 0 \qquad\quad if \;\; \alpha_k = 0, \\ 1 \quad if \;\; \alpha_k = 1,...,l_i - 1, l_i \end{cases}.$$

These new dichotomous subattributes, by definition, satisfy the following order constraint: $\alpha_{k\_1} \leq \alpha_{k\_2} \leq ... \leq \alpha_{k\_l_i}$. In other words, the only allowable combinations of these subattributes are the Guttman patterns: $(\alpha_{k\_1}, \alpha_{k\_2}, ..., \alpha_{k\_l_i}) = (000...00), (000...01), ..., (011...11), (111...11)$.

It can easily be shown that the normal fusion model parameterization applied to these subattributes with the enforced order constraint given above is equivalent to the original parameterization of the ordered polytomous fusion model (op. cit.).

The General Diagnostic Model

This section introduces a GDM (von Davier, 2005; von Davier & Yamamoto, 2004a, 2004b) for dichotomous and polytomous data and ordinal skill levels. The class of diagnostic models is defined by a discrete, multidimensional, latent variable θ, that is, $\theta = (a_1, \ldots a_K)$ with discrete user-defined skill levels $a_k \in \{s_{k1}, \ldots, s_{kl}, \ldots, s_{kL_k}\}$.

In the simplest (and most common) case the skills are dichotomous, that is, the skills will take on only two values $a_k \in \{0,1\}$. In this case, the skill levels are interpreted as mastery (1) versus nonmastery (0) of skill k. Let $\theta = (a_1, .., a_K)$ be a K-dimensional skill profile consisting of K polytomous skill levels $a_k, \quad k = 1, .., K$. Then define the item-specific logits as

$$\log\left[\frac{P(X = x \mid \beta_i, q_i, \gamma_i, a)}{P(X = 0 \mid \beta_i, q_i, \gamma_i, a)}\right] = \beta_{xi} + \sum_{k=1}^{K} \gamma_{xik} h_i(q_{ik}, a_k) \qquad (4)$$

with Q-matrix entries $q_{ik} = \in \{0,1,2,..\}$ and slope parameter $\gamma_{xi} = (\gamma_{xi1}, .., \gamma_{xiK}) \in R^K$. The Q-matrix entries q_{ik} relate item i to skill k and determine whether or not (and to what extent) skill k is required for item i. If skill k is required for item i, then $q_{ik} > 0$, if skill k is not required, then $q_{ik} = 0$.

These $h_i(q_{ik}, a_k) \mapsto R$ are central building blocks of the GDM. The function h_i maps the skill levels a_k and Q-matrix entries q_{ik} to the real numbers. In most cases the same mapping will be adopted for all items, so we may drop the index i. The h mapping defines how the Q-matrix entries and the skill levels interact (See the examples in the next subsection.).

Examples of Skill by Q-Matrix Interactions

One example for a mapping $h_i()$ relates the GDM to discrete multidimensional item response theory (MIRT) models. The choice of h for IRT type models is

$$h_i(q_{ik}, a_k) = q_{ik} a_k \tag{5}$$

which, for $q \in \{0,1\}$, equals

$$h_i(q_{ik}, a_k) = \begin{cases} a_k & for\ q_{ik} = 1 \\ 0 & for\ q_{ik} = 0 \end{cases}.$$

In this case, only the skills k with nonzero Q-matrix entries q_{ik} (the skills required for this item) contribute to the probability of item i. If $q_{ik} = 1$ holds, we have a total contribution of $\gamma_{ik} h(q_{ik}, a_k) = \gamma_{ik} a_k$ for skill k in Equation 4.

The above choice is appropriate for Q-matrices with $0/1$ entries combined with various skill level choices. Skill levels such as $a_k \in \{0,1\}$ or $a_k \in \{-m,\ldots,0,\ldots,+m\}$ may be used with this definition of h as long as the Q-matrix contains only $0/1$ entries.

However, this choice of $h()$ does not work well with Q-matrices that have entries other than $0/1$. This is particularly true if the γ parameters as given in Equation 4 are to be estimated.

In cases with integer or real valued Q-matrices, a useful choice is

$$h(q_{ik}, a_k) = min(q_{ik}, a_k) \tag{6}$$

with $q \in \{0,1,2,\ldots,m\}$ as well as $a \in \{0,1,2,\ldots,m\}$. This coincides with the definition in Equation 5 if $q \in \{0,1\}$ and $a \in \{0,1\}$ but differs in cases using arbitrary skill levels a or Q-matrix entries q.

The rationale of this particular choice of the minimum of q and a is that the GDM may be used for skills assessment where the Q-matrix entries represent a sufficient level for skill k on item i. A higher skill level than q_{ik} will not increase the probability of solving item i, whereas a skill level lower than q_{ik} results in a lower probability of solving item i.[2]

Examples of Skill Level Definitions for Various Models

Assume that the number of skill levels is $S_k = 2$ and choose skill levels $a_k \in \{-1.0, +1.0\}$, or alternatively $a_k \in \{-0.5, +0.5\}$. Note that these skill levels are a priori defined constants and not model parameters. This setting can be easily generalized to polytomous, ordinal skills levels with the number of levels being $S_k = m+1$ and a determination of levels such as $a_k \in \{(0-c), (1-c), \ldots, (m-c)\}$ for some constant c; an obvious choice is $c = m/2$.

Consider a case with just one dimension, for example, $K = 1$, and many levels, say, $S_k = 41$, with levels of a_k being equally spaced (a common, but not a necessary choice) say $a_k \in \{-4.0, \ldots, +4.0\}$. Here, the GDM mimics a unidimensional IRT model, namely the generalized partial credit model (GPCM; Muraki, 1992).

A Logistic Version of the General Diagnostic Model

The log linear formulation of the GDM as given in Equation 4 may be transformed to a form that is more familiar to researchers working with IRT models. The model as introduced above is equivalent to

$$P(X = x \mid \beta_i, q_i, \gamma_i, a) = \frac{\exp\left[\beta_{xi} + \sum_{k=1}^{K} \gamma_{xik} h_i(q_{ik}, a_k)\right]}{1 + \sum_{y=1}^{m_i} \exp\left[\beta_{yi} + \sum_{k=1}^{K} \gamma_{yik} h_i(q_{ik}, a_k)\right]} \qquad (7)$$

with k-dimensional skill profile $a = (a_1, .., a_K)$ and with some necessary restrictions on the $\sum_k \gamma_{xik}$ and the $\sum \beta_{xi}$ to identify the model. This defines the GDM as a general class of skill profile models. von Davier and Yamamoto (2004a) showed that this class of models already contains a compensatory version of the fusion model as well as many common IRT models as special cases. The parameters β_{xi} as well as γ_{xik} may be interpreted as threshold parameters and slope parameters, respectively.

General Diagnostic Models for Partial Credit Data

For a partial credit version of the GDM, choose $h_i(q_{ik}, a_k) = q_{ik} a_k$ together with Q-matrices containing only $0/1$ entries. The resulting model, referred to as pGDM, contains many standard IRT models and their extensions to confirmatory MIRT models using Q-matrices.

Additionally, skill profile models such as multiple classification latent class models (Maris, 1999), located latent class models (Formann, 1985), and a compensatory version of the fusion model (Hartz, Roussos, & Stout, 2002) are special cases for this subset of the GDM**.** For dichotomous and ordinal responses, this member of the GDMs, which may be viewed as a multivariate, discrete, GPCM, $x \in \{0,1,2,..,m_i\}$, is

$$P(X = x \mid \beta_i, a, q_i, \gamma_i) = \frac{\exp\left[\beta_{xi} + \sum_{k=1}^{K} x\gamma_{ik} q_{ik} a_k\right]}{1 + \sum_{y=1}^{m_i} \exp\left[\beta_{yi} + \sum_{k=1}^{K} y\gamma_{ik} q_{ik} a_k\right]} \qquad (8)$$

with k attributes (discrete latent traits) $a = (a_1,..,a_K)$, and a dichotomous design Q-matrix $(q_{ik})_{i=1..I,k=1..K}$. The a_k are discrete scores determined before estimation and can be chosen by the user. These scores are used to assign real numbers to the skill levels, for example, $a(0) = -1.0$ and $a(1) = +1.0$ may be chosen for dichotomous skills. de la Torre and Douglas (2004) estimated the dichotomous version of this model, the linear logistic model (LLM; Maris, 1999; Hagenaars, 1993), using MCMC methods. For ordinal skills with s_k levels, the a_k may be defined using $a(x) = x$ for $x = 0,\ldots,(s_k - 1)$ or $a(0) = -s_k/2,\ldots,a(s_k - 1) = s_k/2$. The parameters of the models as given in Equation 8 can be estimated for dichotomous and polytomous data, as well as for ordinal skills, using the EM algorithm.

The pGDM can be extended also to a mixture distribution IRT model (von Davier & Rost, 2006), which allows the estimation of this class of diagnostic model in different latent classes without prespecifying which observation belongs to which class. This provides the ability to check whether the same kind of skill-by-item relations hold for all the subjects sampled from a particular population. A multiple group version of the pGDM can also be specified and estimated using the algorithm described below. This allows the estimation of diagnostic models, using the GDM framework, that contain partially missing grouping information (similar to the approach described in von Davier & Yamamoto, 2004c). For diagnostic models involving multiple observed groups or multiple unobserved populations (latent classes), parameter constraints can be specified that ensure scale linkages across these populations. Xu and von Davier (2006) presented an application of this approach to data from the National Assessment of Educational Progress (NAEP). NAEP is a government mandated program of monitoring educational progress

in the United States. Many view NAEP, with good reason, as the mother of international large scale educational survey assessments such as PISA—Programme for International Student Assessment , TIMSS—Trends in Mathematics and Science Study, and PIRLS—Progress in International Reading Literacy Study (PIRLS is known as IGLU in Germany).

Estimation and Data Requirements

An implementation of the EM algorithm based on a program for discrete mixture distribution IRT models (von Davier, 2001; von Davier & Yamamoto, 2004c) has been developed. This extended program, called *mdltm*, can be used to estimate parameters of the model as given in Equation 8. The program employs the EM algorithm and provides information about convergence, numbers of required iteration cycles, and descriptive measures of model-data fit and item fit. The program is controlled by a scripting language that is used to describe the data input format and the skill model (i.e., the item-skill combination as given in the Q-matrix, the number of skill levels, skill level scores a_k for each skill, and whether the γ parameters are constrained across items or estimated freely).

The software has been tested with samples of up to 200,000 examinees implementing a two-dimensional IRT model as well as with up to 50,000 examinees and an eight-dimensional dichotomous skill variable $\theta = (a_1,\ldots,a_8)$. Larger numbers of skills very likely will pose problems with identifiability, whether MCMC (in Bayes nets or other approaches) or MML methods are used, unless the number of items per skill variable is sufficiently large. The *mdltm* software allows imposing various types of constraints that may help to achieve identifiability in such cases. Currently, the following skill profile models can be estimated using the software:

- multiple classification latent class models (Maris, 1999), diagnostic models with dichotomous skill variables, a compensatory Fusion/Arpeggio model (sometimes referred to as RUM; Hartz, Roussos, & Stout, 2002);
- direct extensions of these diagnostic models to polytomous response data, and polytomous, ordinal skill levels (von Davier & Yamamoto, 2004a, 2004b; von Davier, 2005), without the need for replacement of ordinal skills by dichotomous sub skills with order constraints;

- unidimensional IRT models such as the RM (Rasch, 1960), the partial credit Rasch model (Masters, 1982), the 2PL IRT (Birnbaum, 1968) model, the generalized partial credit model (Muraki, 1992); and
- other latent structure models, such as located latent class models (Forman 1985; Haberman, 1979); confirmatory multivariate IRT models; and discrete mixture IRT models (von Davier & Rost, 2006), such as polytomous mixed Rasch models (von Davier & Rost, 1995).

The software can read ASCII data files in arbitrary format, and the scripting language used to control the software enables the user to specify which columns represent which variables. The software also handles weighted data, multiple group data (multiple populations), data missing by design (matrix samples) in response variables, and data missing at random in response variables as well as in grouping variables.

The output is divided into a model parameter summary, an estimation summary, and a file containing the scores and attribute classifications of each examinee. This file also contains an accuracy percentage for each subscale as defined by the Q-matrix and the examinee ID code.

Conclusions

A variety of different diagnostic information is used in testing programs. In some programs, subscores or proficiency levels are used rather than diagnostic models. This is often the case for so-called legacy programs mainly created for providing unidimensional scores. For these programs, proficiency levels rather than skill profiles may seem more appropriate, since empirical comparisons often suggest that distinct patterns of response behaviors exist only in a hierarchical sense, namely that students at different levels on the ability scale have difficulties with different subdomains of the test.

Currently, several testing programs use proficiency levels or rule-space-based methods as tools for providing the feedback from test outcomes to students and educators. An example is the PSAT. Other programs such as TOEFL® iBT and survey assessments such as NAEP are currently evaluating models for cognitive diagnosis as a means of providing more meaningful feedback about student (or group) performance on these tests or assessment instruments. Currently, researchers at ETS compare different approaches to cognitive diagnosis on the basis of data from operational test administrations of the TOEFL iBT testing program (Lee & Sawaki,

2006). Von Davier (2005) applied the GDM to TOEFL iBT field test data. Xu and von Davier (2006) studied parameter recovery of the GDM under sparse item matrix designs such as the ones used in large-scale educational assessments including NAEP and PISA.

Using diagnostic models assumes that there are noticeable advantages in modeling examinee performance in a multivariate way. Testing programs have to find a balance between the two competing goals of assessing students' performance in a summative way, while at the same time trying to help students discover their strengths and weaknesses.

Models for cognitive diagnosis that fit into a common framework with IRT and latent class models help to determine whether additional model complexity will pay off in terms of greater accuracy in predicting student responses to the tasks or items in a cognitive assessment. More work needs to be done to derive suitable diagnostics for such complex models, thereby enabling researchers to pinpoint where models fail to predict student responses. Models for cognitive diagnosis expose the fact that we often may have very different ideas of how students interact with tasks. In this framework, those ideas are reflected in either unidimensional views of performance or in hypotheses of how a cognitive domain is structured in multidimensional mastery/nonmastery skills.

How students' skills or abilities are organized is a question that often may not be answered empirically, but the least statistical models can do is provide different ways to predict future performance. This will allow researchers to choose between more or less parsimonious data descriptions that fit the specific questions they wish to answer, given that the descriptions fit the observed data comparably well. This is not to say that we are free to choose, since many applications of test outcomes do not require multiple and potentially correlated measures. If, for example, the question is which students of a certain grade level should be assigned to broad remedial reading or writing classes, a simple classifier seems sufficient, since that does not require a complex model of five to seven interacting (unobserved) student variables that have to be combined in some way to decide whether a student needs special classes or not.

On the other hand, multidimensional skill profiles may serve well in circumstances where enough information is available to measure each of the skills involved sufficiently well. This may be the case when data from multiple assessments are collected, each of which taps into one or more moderately correlated skills and when the goal of the assessment is equally complex, for example, putting together an individualized training program for each student.

References

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Bolt, D. M., & Fu, J. (2004, April). *A polytomous extension of the fusion model and its Bayesian parameter estimation.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Davier, M. von. (2001). WINMIRA 2001. Program and software for Rasch models, mixed Rasch models, latent class analysis and hybrid models [Computer software]. Kiel, Germany: IPN: Institute for Science Education.

Davier, M. von. (2005). *A general diagnostic model applied to language testing data* (ETS RR-05-16). Princeton, NJ: ETS.

Davier, M. von, & Molenaar, I. W. (2003). A person-fit index for polytomous Rasch models, latent class models, and their mixture generalizations. *Psychometrika, 68*(2), 213–228.

Davier, von, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 371–379). New York: Springer.

Davier, M. von, & Rost (2006). Mixture distribution item response models. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 27. Psychometrics.* Amsterdam: Elsevier.

Davier, M. von, Sinharay, S., Oranje, A., & Beaton, A. (2006). *Marginal estimation of population characteristics: Recent developments and future directions.* In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 27. Psychometrics.* Amsterdam: Elsevier.

Davier, M. von, & Yamamoto, K. (2004a, October 21). *A class of models for cognitive diagnosis.* Paper presented at the 4th Spearman Conference, Philadelphia, PA.

Davier, M. von, & Yamamoto, K. (2004b, December 2). *A class of models for cognitive diagnosis—And some notes on estimation.* Paper presented at the ETS Tucker Works seminar, Princeton, NJ.

Davier, M. von, & Yamamoto, K. (2004c). Partially observed mixtures of IRT models: An extension of the generalized partial credit model. *Applied Psychological Measurement, 28*, 389–406.

de la Torre, J., & Douglas, J. A. (2004). Higher order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

DiBello, L. V., & Stout, W. F. (2003). Student profile scoring for formative assessment. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 369–380). New York: Springer.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan, (Eds.), *Cognitively diagnostic assessment* (pp.361–389). Hillsdale, NJ: Erlbaum.

Formann, A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology, 38,* 87-111.

Haberman, S. J. (1979). *Qualitative data analysis: Vol. 2. New developments.* New York: Academic Press.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301-321.

Hagenaars, J. A. (1993). *Loglinear models with latent variables.* Sage: Newbury Park.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* Unpublished doctoral dissertation, University of Illinois, Champaign.

Hartz, S., Roussos, L., & Stout, W. F. (2002). *Skills diagnosis: Theory and practice. User manual for Arpeggio software.* Princeton, NJ: ETS.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272.

Lee, Y-. W., & Sawaki, Y. (2006). *A comparison of diagnostic models based on TOEFL iBT data.* Manuscript submitted for publication.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*(2), 187–212.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Samejima, F. (1995). A cognitive diagnosis method using latent trait models: Competency space approach and its relationship with DiBello and Stout's unified cognitive-psychometric diagnosis model. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 391–410). Hillsdale, NJ: Erlbaum.

Stout, W. F., & Hartz, S. M. (2004). *U.S. Patent No. 6,832,069.* Washington, DC: U.S. Patent and Trademark Office.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345-54.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 10,* 55-73.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: statistical pattern recognition and classification approach. In P. Nichols, S. Chipman, & R. Brennen (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale, NJ: Erlbaum.

Templin, J. (2004). *Generalized linear mixed proficiency models for cognitive diagnosis.* Unpublished doctoral dissertation, University of Illinois, Urbana.

Templin, J., Roussos, L., & Stout, W. F. (2004). *An extension of the current fusion model to treat polytomous attributes.* Unpublished manuscript.

VanEssen, T. (2001, April). *Developing skills, descriptions and steps for improvement on a national standardized test.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Xu, X., & von Davier, M. (2006). *Cognitive diagnosis for NAEP proficiency data* (ETS RR-06-08). Princeton, NJ: ETS.

Yamamoto, K. (1992). HYBIL II software for estimating the HYBRID models and constrained latent class models [Computer software]. Princeton NJ: ETS.

Notes

[1] Note, however, that constraints across classes—skill patterns— may be used to decrease the actual number of required parameters to model the latent skill space.

[2] Assuming fixed skill levels a_l on the remaining skills $l \neq k$ and a slope parameter $\gamma_{ik} > 0$.